中国科学：地球科学 SCIENTIA SINICA Terrae 2026 年 第 56 卷 第 1 期： 117–132
论 文
《中国科学》杂志社
SCIENCE CHINA PRESS

人地系统建模与决策的通用框架初探：多主体深度强化学习及其在公地悲剧中的应用

原世伟，李新*

中国科学院青藏高原研究所，青藏高原地球系统与资源环境国家重点实验室，国家青藏高原科学数据中心，北京 100101
* 通讯作者，E-mail：xinli@itpcas.ac.cn

2025-03-26收稿，2025-08-27修回，2025-10-28接受，2025-12-12网络版发表
国家自然科学基金项目(42301545、42430112)和中国博士后科学基金项目(2023M733606)资助

摘要 人地系统是人类活动与自然过程深度交织的复杂适应系统，随机性、非线性、反馈与涌现等现象交相作用，共同驱动着人类行为与自然动态之间的耦合演化。尽管地球系统科学中的自然系统模型已相对成熟，但对人类系统的建模仍存在明显不足，且缺乏能够有效刻画两者之间复杂交互，支持人地系统建模和决策的通用框架。针对上述问题，本研究提出了一个集成马尔可夫决策过程、多主体建模和深度强化学习的多主体深度强化学习框架，以应对人地系统中建模和决策的挑战。该框架以马尔可夫决策过程四元组为基础：环境状态(表征自然系统)、主体行动(表征人类系统)、状态转移(表征人地系统演化)和奖励机制。通过引入多主体建模来模拟从个人、群体、社区到政府和非政府组织等多层级主体的行为、决策及其复杂交互。深度强化学习则用于解决由此引入的高维状态-动作空间求解难题。最后，以经典的“公地悲剧”为背景，设计了一个由多名依据特定决策规则行动的渔民组成，围绕共同鱼塘资源进行决策的实验案例。结果表明，在仅受经济利益驱动时，渔民们倾向于采取高强度捕捞策略，导致鱼群数量在20个时间步长内迅速由初始的1600单位下降至接近零，重现了经典的“公地悲剧”现象；相比之下，引入可持续性惩罚机制或合作机制能够有效引导渔民调整捕捞策略，表现出更为稳定和适度的捕捞行为，鱼群存量在整个模拟期分别维持在约500和1500单位，实现了资源的可持续发展；此外，通过引入行为参数(贪婪因子)，有效刻画了渔民在捕捞倾向上的异质性，高贪婪因子的渔民表现出激进的捕捞行为，而低贪婪因子的渔民则采取更为保守的策略，并揭示了个体行为差异对系统动态的影响。本研究提出的多主体深度强化学习综合集成框架能够有效捕捉异质主体与环境之间的动态反馈机制与非线性涌现特征，为解析人地系统中多过程、多主体及多活动之间的复杂交互机制提供了有力工具，并为未来开展多尺度系统治理与决策的大规模数值实验奠定了坚实的方法论基础。

关键词 人地系统，马尔可夫决策过程，强化学习，多主体建模，公地悲剧

1 引言

人地系统(Human-Natural Systems)是人类活动与自然过程错综交织的复杂适应系统。要实现对此类系统的有效建模与科学决策，需要深入理解人类行为与自然动态之间受反馈机制、涌现特性、非线性及随机性协同调控的复杂交互过程(Liu等，2007；Steffen等，2020；陈发虎等，2021；刘彦随等，2024)。然而，这些关键特征常引发诸如蝴蝶效应(对初始条件的极端敏感性)和分岔现象(参数变化引起的急剧转变)等复杂行为，最终导致系统混沌的出现(Lorenz，1963；May，1976；Liu等，2007；An，2012；The Nobel Committee for Physics，2021)。由此产生的高度敏感性及不可预测性显著放大了人地系统可持续管理的决策风险，凸显了对先进建模和决策方法的迫切需求(Lansing，2003；Rispoli，2020；Reed等，2022；岳天祥等，2024)。

中文引用格式：原世伟, 李新. 2026. 人地系统建模与决策的通用框架初探：多主体深度强化学习及其在公地悲剧中的应用. 中国科学: 地球科学, 56(1): 117–132, doi: 10.1360/N072025-0092
英文引用格式：Yuan S, Li X. 2026. Generalizing human-natural systems modeling and decision-making: A multi-agent deep reinforcement learning framework and its application to the tragedy of the commons. Science China Earth Sciences, 69(1): 109–124, https://doi.org/10.1007/s11430-025-1738-8

© 2025《中国科学》杂志社

人地系统研究中的若干经典问题，深刻揭示了当前方法论存在的不足。“公地悲剧”作为其中之一，所刻画的由个体自利行为导致共享资源过度开发直至枯竭的现象(Hardin，1968)，凸显了模拟集体行动困境、识别可持续治理路径的必要性。另一个核心问题，“增长的极限”，揭示了地球资源的有限性，以及人口/消费指数式增长对环境产生的深远影响(Meadows等，1972)，催生了对能够平衡经济增长与环境可持续性模型的迫切需求。近年来，Rockström等(2009)提出的“行星边界”框架通过定量界定地球系统的安全运行空间，进一步强化了理解地球承载极限和人类活动潜在越界风险的重要性。同时，对社会临界要素的关注日益提升，该概念强调社会系统中存在关键阈值点，微小扰动即可触发系统性转变，如可再生能源技术的快速普及或环境保护规范的全面崩溃等，且此类临界点常与生物物理临界点之间存在复杂级联效应(Lenton等，2008)。上述经典问题的共同本质在于迫切需要构建一套破译人地系统协同演化动力学及自适应反馈机制的综合框架，以有效应对人类世背景下面临的可持续发展挑战。

然而，现有模型往往无法全面刻画现实世界中人地系统的复杂性，导致其演化过程的模拟与预测面临显著瓶颈(Li Y等，2023)。具体而言，传统方法在表征反馈回路、适应性行为及其协同演化方面存在局限，导致预测结果偏离实际或存在系统性偏差(Farahbakhsh等，2022；Bodin和Chen，2023)。应对人地系统的复杂性挑战，亟需在理论和数学方面协同突破。在理论层面，必须构建能够深度阐释系统基本运行机制的综合框架，系统整合反馈机制、适应性行为和涌现特性等关键要素(Levin等，2013)。但目前仍缺乏一个能够有效指导建模的统一数学框架。该框架不仅需要能够处理高维空间和非线性动力学，还需要解决系统随机性以及适应性演化问题。缺乏这样的数学框架，将严重阻碍对人地系统复杂适应相互作用的有效模拟与预测，进而限制了对各种管理策略与政策干预措施潜在效应的探索与评估(Holling，1973)。

针对上述需求，以下几种方法论在人地系统复杂性研究中展现出独特优势。马尔可夫决策过程(Markov Decision Process)为不确定性和随机性环境下的序贯决策问题提供了结构化的概率框架(Bellman，1957a，1966；Puterman，1990)，其策略的长期优化能力对于环境和社会的可持续发展管理至关重要。基于马尔可夫决策过程框架，强化学习(Reinforcement Learning)显著增强了对高维复杂决策问题的处理能力(Sutton和Barto，2018)。其核心在于主体能够通过试错交互机制学习最优策略，在动态变化的环境中不断适应与优化。这种学习机制深度契合人地相互作用过程中的动态演化本质。多主体建模(Agent-Based Modeling)则从根本上弥补了上述方法的建模视角局限，通过定义异质性主体的规则与行为，自下而上地模拟微观交互引发的集体涌现现象(Bonabeau，2002；Grimm等，2005；Schulze等，2017；Yuan等，2021)，有效刻画了传统方法难以处理的动态适应性和涌现现象。然而，要深入理解并有效管理人地系统，亟需构建一个能有机融合上述方法优势的统一框架，既具备自下而上捕捉复杂适应性行为与涌现现象的能力，又兼具自上而下进行长期最优策略优化与决策的能力。

综上，人类世背景下的人地系统是一个动态、开放且复杂的巨系统，其中不仅蕴含着利益观、世界观和价值观的深层冲突(钱学森，1991)，同时面临着公地悲剧与可持续发展等严峻挑战，这样的系统在复杂性科学领域也被称为复杂适应系统。该类系统的演化与稳定性核心取决于多主体(或子系统)的局部行为决策及其向系统全局的反馈传导(Li等，2022)。当前研究已深化为对系统复杂性的追踪、溯源、解析、建模和决策支持的综合性探索，这不仅需要对人地系统概念的持续精炼，更迫切需要发展更加可解释可操作的方法论。为此，本研究提出并构建了一个通用的多主体深度强化学习(Multi-Agent Deep Reinforcement Learning)框架。该框架的核心创新在于实现了马尔可夫决策过程、强化学习和多主体建模三种方法的耦合集成，将马尔可夫决策过程/强化学习自上而下的决策优化与多主体建模自下而上的行为建模有机统一，为人地系统建模和决策提供了更加全面的视角。最后，以经典的“公地悲剧”为背景，设计了一个由多名依据特定决策规则行动的渔民组成，围绕共同鱼塘资源进行决策的实验案例，以验证所提框架的有效性和实用性。

2　理论回顾与梳理

尽管在过去几十年中，与人相关的理论与方法框架取得了显著进展(Li等，2018，2022)，然而由于缺乏普适性的整合方法，对人地系统的全面认知及有效决策仍面临巨大挑战。如图1所示，人的决策与学习理论方法经历了从认识论探索到多主体深度强化学习方法论突破的若干重要阶段，其核心理论方法概述如下。

2.1　哲学基础：人的科学

将人类行为与决策纳入自然系统建模框架的思想源起于启蒙运动时期，其哲学根基可追溯至大卫·休谟对“人的科学”(Science of Man)的开创性探索(Hume，

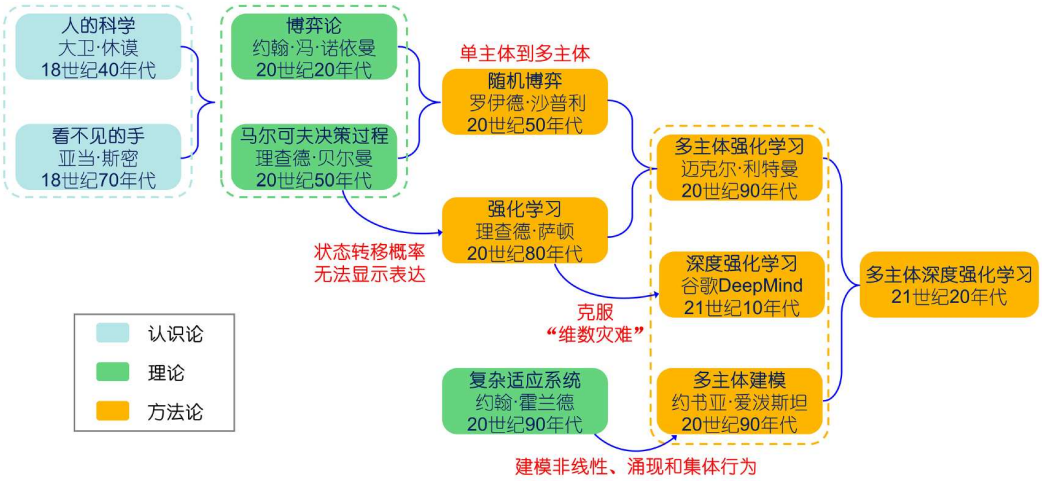


图1 决策与学习理论方法的发展演进历程。主要包含三个阶段：认识论奠基阶段(休谟的经验主义和斯密的理性主义)、理论发展阶段(博弈论和马尔可夫决策过程)以及方法论创新阶段(随机博弈、强化学习、深度强化学习等)。多主体深度强化学习的出现代表了计算智能与类人推理相结合的前沿领域

1739；Berry，2007)。休谟主张科学研究应当超越自然现象的范畴，将人类行为和决策纳入其中(Hume，1739)。这一洞见从经验哲学层面确立了人类作为自然系统有机组分的地位，为研究人类行为提供了哲学基础。从这一思想脉络看，当代人地系统中对人类决策行为的研究正是“人的科学”的现代延展，正如休谟所指出的，“凡属重要问题，其答案都包括在人的科学中”(There is no question of importance, whose decision is not comprised in the science of man)(Hume，1739)。

亚当·斯密的“看不见的手”(Invisible Hand)理论为休谟的框架提供了重要的社会经济学延拓(Smith，1776)，阐述了在特定道德和社会规范约束下，个体层面看似自利的行为能够在系统层面自发涌现出促进整体福利的宏观秩序。其核心价值在于揭示了复杂社会系统如何从微观个体行为动态中涌现，深刻呼应了当代复杂适应系统理论中对适应性行为与系统涌现关系的关注。休谟与斯密的奠基性工作共同确立了“人的科学”的基本方向，标志着对人类行为进行系统性认识论探索的开端。

然而需要指出的是，休谟与斯密的贡献虽具有划时代的意义并奠定了关键性的概念基石，但主要停留于哲学认识论层面。他们尚未发展出解析动态复杂系统内人类交互行为的系统化数学方法与工具。

2.2 现代决策理论和马尔科夫决策过程

基于休谟与斯密所奠定的哲学基础，现代决策理论在20世纪应运而生。其兴起标志着人类行为和决策研究从认识论层面向实用方法论转变。这一演变的核心在于博弈论的创立与发展。该理论于20世纪20年代由约翰·冯·诺依曼和奥斯卡·摩根斯坦开创(Von Neumann，1928；Von Neumann和Morgenstern，1944)。博弈论形式化描述了理性主体间的策略互动，即每个拥有自身偏好与目标的个体，其决策行为相互制约并可能导致出乎意料的结果。这一框架为解析战略互动、合作/非合作行为模式，以及在资源管理和冲突等复杂情境下的决策动态提供了强大的理论根基。

与此同时，在20世纪50年代，理查德·贝尔曼提出了马尔可夫决策过程，为不确定性环境中的序贯决策问题建立了严格的数学框架(Bellman，1954，1957a，1957b，1966)。马尔可夫决策过程为单个主体在动态环境中的决策提供了一种结构化方法，其核心原理在于每步决策仅依赖于当前系统状态以及向未来状态的概率转移关系。贝尔曼的开创性贡献在于严谨刻画了动态随机环境中的最优决策过程，其所提出的贝尔曼方程通过递归形式精确表征了序贯决策最优策略的性质。马尔可夫决策过程不仅为理解人类在动态不确定环境下的决策行为提供了坚实的数学基础，更为后续强化学习等自适应决策算法的发展铺设了核心理论基石。

值得注意的是，博弈论主要关注多主体间交互的战略均衡，而经典的马尔可夫决策过程框架则侧重于单主体与环境之间的动态决策问题。两者的概念融合将马尔可夫决策过程框架扩展到多主体共享环境的情境。在此扩展框架(被称为随机博弈)中，每个主体的决策与奖励不仅取决于自身行为，更显著依赖于其他主体的策略选择以及环境交互的概率特性。这种拓展使实现对多主体系统中复杂动态交互的建模成为可能。

现代决策理论、博弈论与马尔可夫决策过程为20世

纪80–90年代强化学习的兴起奠定了重要基础(Minsky，1961；Skinner，1965；Werbos，1977；Sutton，1984；Watkins，1989)。强化学习专注于主体(人类或者人工)在环境交互中的自适应决策过程(Sutton和Barto，2018)。植根于行为心理学，强化学习揭示了主体如何在试错探索中逐步习得最优决策策略，该过程与人类通过环境反馈不断调适行为模式的认知模式在本质上具有高度相似性。这一特质为决策理论和博弈论在自主系统、智能机器人和人地系统等复杂动态环境中的实际应用提供了核心算法引擎与计算框架。

这一发展阶段标志着从认识论理解(理解人类行为)向方法论构建(建模和预测人类行为)的重要转变。强化学习正是这种转变的典型代表，为不确定性动态环境中决策问题的求解和优化提供了强大的算法方法。这些理论与算法进步显著提升了对人地系统复杂性挑战进行建模和响应的能力，实现了理论洞察与计算实践的结合。

2.3　复杂适应系统和多主体建模

复杂适应系统(Complex Adaptive Systems)作为一门融合了生物学、生态学、经济学和系统科学的交叉学科，兴起于20世纪下半叶(Holland，1992；Lansing，2003)。该领域的奠基性工作主要发轫于20世纪80年代的圣塔菲研究所，约翰·霍兰德、穆雷·盖尔曼和斯图尔特·考夫曼等学者对复杂交互系统的内在动力学机制展开了系统探索。复杂适应系统区别于传统复杂系统的本质在于其深刻的适应性特征，尤其是对环境变化的响应与调整能力。这种适应性源于主体与环境之间持续的互馈循环：主体在动态变化的环境中学习并作出响应，其响应行为又反过来塑造环境演化轨迹。这种强反馈特性使得对复杂适应系统的解析远较传统的“适定”问题或“病态”问题更具挑战性。尤需关注的是，由于广泛存在的非线性相互作用和涌现现象，复杂适应系统往往表现出根本性的不可预测性。即使已知系统的精确初始条件与演化规则，对其进行长期预测通常仍不可行。这种不可预测性在社会维度(涉及多样化的利益诉求、文化背景及价值观冲突)尤为突出。诸多现实难题本质上都是复杂适应系统挑战的具体表现：“公地悲剧”反映了个人理性与集体福祉的结构性冲突；“增长的极限”揭示了社会经济发展与资源约束之间的紧张关系；“可持续发展目标”则试图在经济增长、社会公平和生态保护之间寻求平衡。有效应对这些挑战需要同时考虑主体的适应性决策过程、多元利益相关者的需求，以及系统在不同时空尺度上的动态演化特性。

涌现是复杂适应系统的另一关键特质，指由微观层面主体互动自发形成的宏观层面系统行为(Holland，1995，2002)。这些宏观行为通常具有不可预测性和非线性，对传统基于线性或均衡假设的数学建模方法构成显著障碍。多主体建模正是在此背景下发展成的专门用于模拟复杂适应系统的计算方法(Epstein和Axtell，1996；Bankes，2002；Bonabeau，2002；Farmer和Foley，2009)。多主体建模能够显式模拟具有异质行为和适应策略的主体，每个主体遵循其状态、环境交互和主体间互动的特定决策规则，这使得其成为解析“微观主体行为如何导致宏观涌现”机制的有力工具(原世伟等，2021)。

与此同时，多主体建模对强化学习在复杂动态系统中的应用提供了支持(Bone和Dragićević，2010；Lee等，2017；Sert等，2020；Kedir等，2022；Ale Ebrahim Dehkordi等，2023)。多主体建模构建高度灵活的虚拟实验环境，使得强化学习能够有效地模拟并研究多样化主体(如个体、组织、生态单元)在复杂环境中的适应性决策及其交互动态。这种结合显著增强了对人地复杂适应系统进行建模、理解和管理的模拟能力。

2.4　深度强化学习和人工智能突破

尽管强化学习在诸多领域展现出显著潜力，但其向大规模现实场景(尤其是人地系统等高维复杂系统)的扩展仍面临核心挑战(Li X等，2023)，其中一个核心瓶颈便是维数灾难(Curse of Dimensionality)(Hernandez-Leal等，2019)。随着状态空间与动作空间维度的增加，传统强化学习算法的计算复杂度呈指数级增长，导致其难以有效处理现实世界中普遍存在的高维决策问题。同时，环境复杂性的加剧显著提升了从交互中获取有效反馈的难度，进一步阻碍了主体进行高效探索和策略优化的能力。

深度强化学习的突破在很大程度上克服了上述障碍，这得益于基础架构的革新(Mnih等，2015)以及在博弈领域(如围棋、国际象棋)的里程碑式应用(如AlphaGo、AlphaZero)(Silver等，2017，2018)。深度强化学习的核心创新在于融合了深度学习的表征学习能力与传统强化学习的序贯决策机制(Wang等，2022)。一方面，深度神经网络能够自动学习高维输入(如图像、文本和音频)的低维表征(特征)(Arulkumaran等，2017)，极大地降低了处理复杂状态-动作映射的计算负担。另一方面，将其与蒙特卡罗树搜索(Monte Carlo Tree Search)等先进搜索算法相结合，实现了对庞大决策空间的高效探索。蒙特卡罗树搜索算法通过选择性采样和评估高潜力动作序列，显著减少了盲目搜索的成本。更为关键的是，两者形成了强大的学习循环：神经网络为树搜索提供高效的策略与

价值引导，而树搜索生成的高质量数据则持续驱动神经网络的性能提升。此外，深度强化学习利用深度神经网络作为强大的函数逼近器，能够直接学习复杂的价值函数和策略参数化表示，从根本上规避了传统表格或非参数方法中必须显式枚举状态-动作对的繁琐过程(Bengio等，2013)。这种集成赋予了深度强化学习处理复杂策略空间的能力，既实现了卓越的样本效率，也具备了优异的策略泛化性。这些特性显著降低了计算复杂性，使得主体能够在复杂的动态环境中进行高效学习与决策优化。

通过上述创新，深度强化学习为解决复杂高维决策任务提供了可行路径，为将强化学习应用于人地系统等大规模现实问题的建模与决策架设了关键的桥梁。

3 方法

3.1 概念框架与核心组成

基于上述理论方法体系，本研究提出了一种集成马尔可夫决策过程、多主体建模及深度强化学习前沿技术的多主体深度强化学习框架(见图2)。

该框架通过三重机制整合实现对人地系统的协同建模与解析。马尔可夫决策过程为人地系统提供了形式化建模框架，明确定义了其核心要素以及人地交互的动态反馈回路。这种结构化表达使人地系统得以表征为具有内在反馈调节机制的非平衡态动力学系统，从而更本质地反映其现实复杂性。多主体建模实现了微观-宏观机制连接，通过显式模拟异质主体的适应性决策规则及其与环境/其他主体的交互，有效解析微观层面个体/群体行为如何通过非线性相互作用驱动宏观涌现现象。深度强化学习克服了高维优化瓶颈，通过深度神经网络对高维状态/动作空间进行表征学习，并结合贝尔曼方程驱动的策略优化机制，可以在无需完备先验知识的条件下，实现对复杂策略空间的高效探索与优化求解。

通过同时捕捉人地系统的动态演化、主体适应性及涌现特性，该框架为解析复杂人地系统提供了兼具理论性与可操作性的集成工具，显著增强了对人地系统的建模与决策能力。

3.2 马尔可夫决策过程和贝尔曼方程

马尔可夫决策过程为解决不确定性环境下的序贯决策问题建立了严谨的数学框架，通过系统评估策略空间并形式化求解长期累积收益最大化的最优策略，将涉及人地互动的复杂决策研究从哲学认识论层面，推进至可计算、可操作的方法论层面。

马尔可夫决策过程的核心在于其形式化的四元组结构：状态、动作、状态转移概率和奖励(S、A、P、R)(图2)。状态(S)表征系统的关键环境配置，包括水文特征(如流域蓄水量)、生态指标(如生物多样性指数)以及社会经济参数(如区域碳排放水平)等多维动态变量。动

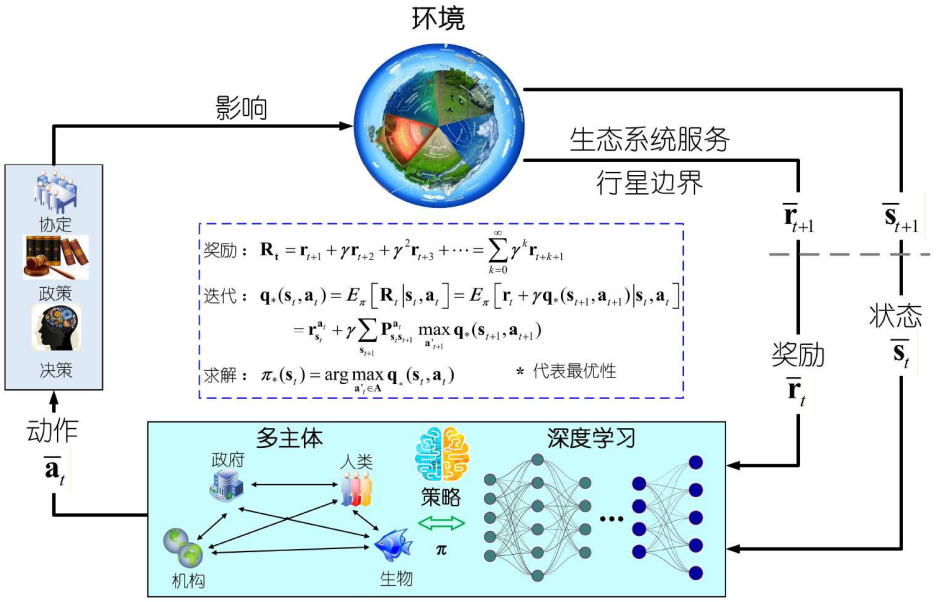


图2 多主体深度强化学习框架。其核心架构基于多主体马尔可夫决策过程，其中主体(如政府、机构、人类和生物)通过协同决策与动态环境状态交互，系统基于交互结果生成实时奖励信号。这些信号驱动深度神经网络对主体策略进行持续优化，形成“环境交互-奖励反馈-策略更新”的自适应学习闭环。该反馈机制使多主体系统能够通过试错学习渐进式提升整体决策效能。贝尔曼方程为该框架提供了数学基础，框中所示的方程代表了贝尔曼方程的最优形式

作(A)定义了决策主体可行的干预措施，例如调整碳税以控制排放，或扩大保护区以遏制森林砍伐等。状态转移概率(P)以条件概率形式量化了主体决策行为对环境状态的影响，例如在调整碳税(动作)后，系统有80%的概率从当前排放状态转变到低碳排放状态。奖励(R)作为即时反馈机制，通过平衡社会生态效益(如生物多样性增益)与经济成本(包括直接经济损失与政策执行投入)，实现多目标优化决策的量化权衡。

(1)主体和环境

主体作为治理决策的执行者，通过感知环境状态、执行干预动作、接收奖励反馈的闭环机制与环境持续交互。环境则表征决策系统所处的动态场景，其状态转移受到主体行为的深刻影响。

(2)状态(S)

状态空间代表系统或环境可能存在的所有配置集合。每个状态$s\in S$都包含了特定时刻的关键决策信息。根据系统复杂度，状态空间可以是有限或无限的，离散或连续的。

(3)动作(A)

在每个决策状态下，主体可执行的干预措施集合构成动作空间，表示为$a\in A(s)$。其同样可以是有限或无限的，离散或连续的。动作选择直接影响环境状态的动态演化，同时决定即时奖励的反馈强度。

(4)状态转移概率(P)

状态转移概率$P\left(s_{t+1}\mid s_t,a_t\right)$描述了环境在给定动作条件$a_t$下从当前时刻状态$s_t$转移到下一时刻状态$s_{t+1}$的概率。其反映了环境中存在的不确定性，以及马尔可夫性，即未来状态仅依赖于当前状态和动作，而不依赖于之前的状态。然而，在大多数现实世界的人地系统中，状态转移概率无法显示表达，必须通过仿真模拟、历史数据或强化学习算法进行估计。

(5)奖励(R)

奖励$R(s,a)$量化了在状态s下采取动作a后所获得的即时收益。其为评估动作的预期效用提供关键依据，并作为决策者长期策略优化的基础目标，即在时序决策中最大化累积奖励。

马尔可夫决策过程的最终目标是找到最优策略$\pi^*:S\rightarrow A$，该策略能够为每个状态指定最大化长期收益的动作选择规则。累积奖励通常表示为价值函数$V^{\pi}(s)$，即从状态s开始并遵循策略π的预期回报：

$$V^{\pi}(s_t)=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,\pi(s_t))\mid s_0\right] \tag{1}$$

式中，$\gamma\in[0,1)$为折扣因子，决定了未来奖励相对于即时奖励的重要性。

另一个关键概念是动作价值函数$Q^{\pi}(s,a)$，它代表在状态s下采取行动a然后遵循策略π的预期累积奖励：

$$Q^{\pi}(s_t,a_t)=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)\mid s_0,a_0\right] \tag{2}$$

最优策略π^*使价值函数或等价地使动作价值函数最大化，进而得到贝尔曼最优性方程：

$$V^*(s_t)=\max_{a'_t\in A}\left[R(s_t,a'_t)+\gamma\sum_{s_{t+1}}P\left(s_{t+1}\mid s_t,a'_t\right)V^*(s_{t+1})\right] \tag{3}$$

$$Q^*(s_t,a_t)=R(s_t,a_t)+\gamma\sum_{s_{t+1}\in S}P\left(s_{t+1}\mid s_t,a_t\right)\max_{a'_{t+1}}Q^*(s_{t+1},a'_{t+1}) \tag{4}$$

$$Q^*(s_t,a_t)=\sum_{s_{t+1}\in S}P\left(s_{t+1}\mid s_t,a_t\right)\left[R(s_t,a_t,s_{t+1})+\gamma\max_{a'_{t+1}}Q^*(s_{t+1},a'_{t+1})\right] \tag{5}$$

式(4)和式(5)的差异在于，前者中的奖励函数仅依赖于当前状态和动作，而后者中的奖励函数进一步与状态转移结果，即执行动作后的后继状态相关。上述方程构成了包括价值迭代与策略迭代在内的最优策略求解算法的数学基础。

3.3　多主体建模

多主体建模是一种通过模拟自主主体在特定环境中的交互行为与决策过程，以研究复杂系统涌现现象的计算建模方法(原世伟等，2021)。在多主体深度强化学习框架中，多主体建模从以下方面弥补了传统马尔可夫决策过程的局限性。

(1)主体异质性

传统的马尔可夫决策过程通常假设单一决策者或同质主体，而多主体建模则显式刻画具有差异化属性、偏好、约束条件与决策规则的异质主体。这种异质性对真实地表征人地系统中的多元利益相关者至关重要，其知识体系、价值取向、社会规范及目标函数的差异将显著影响系统动态演化。

(2)自适应决策规则

多主体建模通过规则演化、模仿学习与文化传承等适应机制，扩展了传统马尔可夫决策过程的学习框架。此类机制更精准地反映了人类主体在复杂系统中动态调整决策行为的特征。

(3)社会互动

多主体建模在基础的主体-环境交互模式上，融入了社会结构、通信网络与主体间交互机制。对社会维度的解析使模型能够表征现实世界中塑造集体结果的关键因

素(如合作、竞争、模仿及冲突)。

(4)涌现

传统马尔可夫决策过程依赖预定义的状态转移概率，而多主体建模则通过主体互动内生产生状态转移规则。这一本质差异使多主体深度强化学习框架能够捕获由微观层面主体互动产生的，但无法直接编码到传统转移矩阵中的涌现现象(如临界点、级联效应和自组织模式)。

集成多主体建模的上述功能特性，多主体深度强化学习框架能够有效刻画传统马尔可夫决策过程无法捕捉到的人地系统复杂性、自适应性及涌现特征。

4 “公地悲剧”案例

本节采用经典的“公地悲剧”案例来验证所提出多主体深度强化学习框架。“公地悲剧”这一概念阐释了共享资源管理的核心困境，即个体理性决策与集体资源可持续性之间的根本矛盾(Hardin，1968；Ostrom，1990)。本案例旨在通过模拟有限资源下多个主体的决策及互动，刻画系统涌现行为并识别促进可持续资源管理的策略，为公地资源治理提供更深入的洞察和更稳健的解决方案。

4.1 实验设计

在渔业背景下，“公地悲剧”概念显得尤为重要，过度捕捞不仅可能导致鱼类种群崩溃，还会威胁资源的可持续性以及依赖该资源的生计(Sethi等，2005；Rudd和Branch，2017；Memarzadeh等，2019)。本研究以一个经典的渔业管理问题为例进行实验来验证所提出的框架(图3)。

实验中，多名依据特定决策规则行动的渔民(主体)围绕一个共享鱼塘资源(环境)进行互动。鱼群的状态转移和渔民的决策过程可建模为一个马尔可夫决策过程(Williams，2011；Ju等，2025)。这本质上反映了经典的公地资源管理问题：渔民的集体捕捞策略共同驱动着鱼群的动态变化。

4.1.1 环境：鱼塘

状态：主要包括鱼群数量S_t、天气条件和其他环境因素。鱼群数量的动态变化由逻辑增长模型描述：

$$F_{t+1}=F_t+\alpha F_t\left(1-\frac{F_t}{E}\right)-\sum C_{i,t} \tag{6}$$

式中，F_t代表时刻t的鱼群数量，α是鱼群的内生增长率，E是环境承载能力。本案例将其设定为α=0.05和E=3000。$C_{i,t}$是渔民i从时刻t到t+1的捕鱼量。

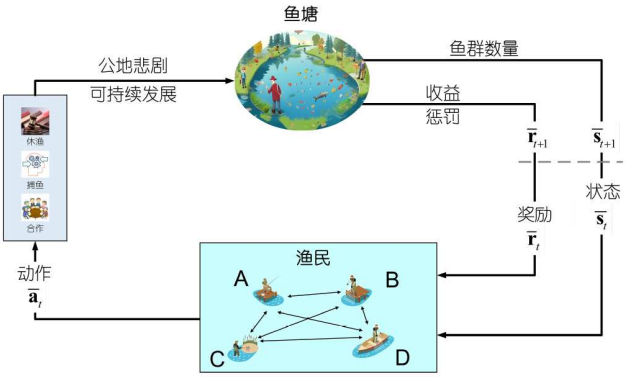


图3 鱼塘案例概念框架。四名渔民根据鱼群的当前状态及所获奖励选择动作。其共同行动将影响鱼塘环境，进而决定鱼群下一阶段的状态，并产生新的奖励(包含收入及潜在惩罚)。此反馈循环促使渔民随时间逐步调整策略，由此可能导致“公地悲剧”或实现“可持续发展”两种结局

4.1.2 主体：渔民

行为参数和规则：每个渔民被赋予一个表征行为特征的贪婪系数g，该系数取值范围为0–1，用于刻画其捕鱼策略的激进程度。贪婪系数较高的渔民表现出激进的捕鱼行为特征，这类渔民更倾向于追求短期经济利益的最大化，对过度捕捞可能导致的资源枯竭问题关注度相对较低。与之相对，贪婪系数较低的渔民则展现出保守的捕鱼行为模式，他们更注重采用可持续的捕鱼策略，通过控制捕捞强度、遵循资源保护准则等方式，致力于为后代保存充足的渔业资源。基于此参数，每个渔民的具体捕鱼行为C_i可通过下式描述：

$$C_i=g_i\times a_t\times SC \tag{7}$$

式中，C_i表示渔民i捕获的鱼群数量，g_i是渔民i的贪婪因子，反映渔民的捕鱼倾向。a_t是捕鱼动作系数，反映了渔民在时刻t的行动强度，设定为0–10之间的整数，以便于模型的离散化处理和计算实现。SC是一个缩放因子(SC=8)，其作用是将渔民的抽象捕鱼动作转换为对应的实际捕鱼量。这种定义方式在资源经济学和行为生态学中具有广泛的理论基础，充分体现了个体主体基于自身动机和外部环境约束进行理性决策的行为特征。

决策：渔民的决策受到鱼群状态(如种群数量与增长率)和个体收益机制(包括捕获量与经济激励)的共同影响。他们基于历史经验动态调整策略，在合作(如资源共享)、监督(如自我调节或监控)或竞争(如个人收益最大化)之间进行转换。这种自适应行为策略塑造了差异化的资源开发模式，渔民在持续优化个体收益的同时，会评估其行为对群体资源的整体影响，并在行动中不断适应环境变化。

4.1.3　情景设置

本研究设计了三种情景来评估不同激励机制对渔民行为和鱼群动态的影响。

(1)基准情景(纯经济奖励)

该情景模拟渔民仅以个人短期收益最大化为目标的行为决策模式。在此设定下，渔民仅考虑个体经济利益，忽略其行为对鱼类种群的长期可持续性影响。由于缺乏集体约束与长远考量，该情景将不可避免地导致“公地悲剧”——渔民个体间竞争性捕捞行为引发种群快速衰退直至枯竭，呈现典型的不可持续发展模式。

(2)可持续性惩罚情景

该情景旨在通过引入惩罚性机制来约束过度捕捞行为。当鱼类种群数量降至预设的可持续性阈值时，系统将对继续超量捕捞的行为施加惩罚，核心目标是延缓资源枯竭速度。惩罚机制的有效性高度依赖于惩罚力度与渔民响应方式。预计能在一定程度上减缓资源耗竭进程，但其最终能否阻止种群崩溃，取决于渔民对惩罚机制的敏感性及其行为调整程度。

(3)合作激励情景

该情景旨在引导渔民同时关注个体收益与集体可持续性目标。通过设计奖励机制，鼓励合作行为并引导个体捕捞量向社群平均水平靠拢，从而促进有利于鱼类种群长期健康发展的捕捞行为。预计渔民将在个体收益与集体可持续性之间寻求更优的均衡点，最终形成更具韧性的可持续性捕捞策略，促使鱼类种群维持在一个更为健康、稳定的水平。

4.2　奖励函数

每位渔民所获得的奖励是其捕捞量的函数，并可根据过度捕捞所施加的惩罚以及潜在合作所带来的激励进行调整。奖励函数的形式具有多样性，具体取决于所模拟的情境设定。在本研究中，设计了三种不同的奖励机制，以模拟和分析渔民在决策过程中的行为特征。每种奖励结构代表了从个体主义到集体行动的不同行为倾向。下文将详细阐述各奖励机制的形式、对应的数学表达式以及相关参数的含义。

(1)单纯经济奖励

$$r_t = \sum C_{i,t} \times \delta_r \tag{8}$$

式中，$C_{i,t}$表示渔民i在时刻t的捕鱼量，δ_r是捕获量的奖励系数，设定为0.05。该奖励函数仅以渔民的个体捕捞量为基础对其进行奖励，未考虑其行为对鱼类种群整体的影响，也未纳入其他渔民行动所产生的效应。

(2)可持续性惩罚

$$r_t = C_i \times \delta_r - \text{Penalty}(F_t) \tag{9}$$

式中，Penalty(F_t)是与鱼群数量F_t相关的惩罚函数。在本研究中，当F_t≤100时施加200的惩罚。该奖励函数通过在鱼类种群数量下降至某一临界阈值以下时对渔民施加惩罚，从而引入可持续性约束，以抑制过度捕捞行为。

(3)合作激励奖励

$$r_t = \sum C_i \times \delta_r - \text{Penalty}(F_t) - \frac{\sum C_i}{n} \times \delta_c \tag{10}$$

式中，n是渔民数量，δ_c是合作系数，设定为0.6。该奖励函数根据所有渔民的平均捕捞量对个体奖励进行调整，从而在渔民之间引入合作机制，旨在通过激励集体行动，减少过度捕捞行为，促进渔业资源的可持续管理。

4.3　强化学习算法

作为集成框架的基础概念验证，本案例选用Q-learning算法。其简洁的结构以及在离散状态-行动空间中具备的高可解释性，使其成为实现本研究主要目标的理想方法，能够清晰揭示主体决策与环境状态之间的反馈动态。尽管Q-learning算法能够有效处理本案例的离散问题，但它难以应对现实世界人地系统面临的高维或连续状态空间挑战。为此，所提框架的核心设计采用了深度强化学习(如深度Q网络)，旨在克服这种“维数灾难”。深度强化学习利用深度神经网络近似Q函数来实现这一点，使框架能够泛化至广阔的状态空间(如空间数据、连续变量)，并支持更复杂的主体决策(Mnih等，2015；Wang等，2022)。因此，Q-learning算法为本次初步探索提供了稳健且便于理解的方法基础，而深度强化学习则为框架向更现实、更复杂场景的扩展确定了明确路径。

Q-learning算法是最常用且最基础的强化学习算法之一，属于基于价值的方法类别，是一种离线策略算法，这意味着其最优策略的学习过程能够独立于主体的当前行为策略。Q-learning算法的核心思想是通过迭代更新动作价值函数，$Q(s_t, a_t)$，使其逐渐逼近真实的最优动作价值函数，$Q^*(s_t, a_t)$，从而得到最优策略。Q-learning算法的更新规则如下：

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma \max_{a'_{t+1}} Q(s_{t+1}, a'_{t+1}) - Q(s_t, a_t)\right] \tag{11}$$

式中，α是学习率；r_{t+1}是执行动作a_t后获得的奖励；γ是折扣因子，它决定了未来奖励的重要性；s_{t+1}是下一时刻状态；a'_{t+1}是在下一个状态时使Q值最大化的动作。

4.4 结果

4.4.1 人类动机对鱼群的影响

为深入探讨不同管理策略对鱼类种群动态的影响，本研究设计了两种情景：情景一以最大化经济收益为目标，情景二则侧重于实现可持续发展目标。

情景一：经济收益导向

在该情景下，奖励结构被设置为优先考虑短期经济回报。如图4所示，四名渔民(A、B、C、D)在开展捕鱼活动时，唯一的目标是最大化自身经济收益。这一设定模拟了缺乏监管机制或过度捕捞惩罚的情境，呈现出典型的“公地悲剧”现象。如图顶部所示，各渔民的捕捞强度虽略有差异，但总体上均维持在较高水平，导致鱼类资源被系统性地过度开发。小提琴图展示了各渔民在决策过程中行动分布的密度，可以看出，其行为高度集中于动作空间的上限区间(7–10)，表明所有个体均倾向于采取高强度捕捞策略。

图4的右下部分描绘了鱼群的动态变化趋势，如绿色曲线(基于逻辑增长模型)所示。由于渔民普遍关注短期收益最大化，鱼群数量迅速减少，最终趋于崩溃。该情景清晰地揭示了在缺乏外部干预的情况下，个体理性行为如何演变为集体非理性的资源耗竭结果，突显了无管制经济激励可能带来的严重生态后果。

情景二：可持续发展导向

相比之下，情景二在奖励结构中引入了可持续性惩罚机制，以防止鱼类资源的过度开发。具体而言，当鱼群数量下降至预设的临界阈值(本研究设定为100单位)以下时，系统会对渔民施加惩罚。这一机制的引入显著改变了渔民的行为模式。在可持续性导向的激励机制引导下，四名渔民的捕捞行为趋于温和。如图5所示，与情景一相比，各渔民的平均捕捞强度显著降低。渔民C表现出最低的平均捕鱼强度(Mean=1.73，Std=2.51)，其次是渔民B(Mean=2.14，Std=3.17)、渔民A(Mean=2.20，Std=3.12)和渔民D(Mean=4.88，Std=2.77)。小提琴图展示了各渔民在决策过程中行动值的分布情况，可以看出，绝大多数行动集中在1–5的低强度区间，与情景一中普遍存在的高强度捕捞行为形成鲜明对比。

这种更加谨慎的捕鱼行为进一步反映在鱼群的动态变化中。不同于情景一中鱼类资源迅速枯竭的趋势，情景二下的鱼群数量在初期有所下降，但随后趋于稳定，在20个时间步长后稳定在大约500单位。该结果展示了如何通过精心设计的激励机制，促使个体在追求经济利益的同时兼顾资源的长期可持续性，从而实现个体与集体目标的协调统一，表明了将可持续性导向政策纳入管理框架的有效性。

两种情景的对比结果突出了政策设计在资源管理中的核心作用。忽视可持续性的激励机制容易引发过度捕捞，导致资源迅速枯竭；而引入针对资源下降的惩罚机制，则能够有效引导渔民采取更为负责任的捕捞行为，从而维持鱼群的长期稳定。这一发现为寻求在经济发展与生态保护之间实现平衡的政策制定者和渔业管理者提供了重要的实践启示(Costello等，2020)。

4.4.2 行为参数对渔民决策制定的影响

除了奖励结构之外，个体的行为特征在决策过程中也发挥着重要影响。为了刻画渔民在行为上的差异，本研究引入了一个行为参数，即贪婪因子g，用于表征个体对即时收益的偏好程度。该参数的取值范围为0–1，其中0表示个体采取保守、风险规避的策略，而1则代表个体倾向于激进、追求短期高回报的风险偏好行为。

图6展示了具有不同贪婪因子的渔民在捕鱼行为上的差异(例如，渔民B的g=0.9，渔民C的g=0.3)。结果显示，贪婪因子较高的渔民倾向于采取更频繁且更具侵略性的捕捞策略，其捕鱼量表现出更高的波动性。以渔民B为例，其平均捕捞强度和行为变异性均显著高于贪婪因子较低的渔民C，后者则展现出更为保守且稳定的捕捞模式。小提琴图进一步揭示了不同渔民之间在捕捞强度上的显著异质性。这种行为模式的差异化分布验证了模型能够有效利用行为参数刻画渔民个体间多样化决策行为的能力。

从系统层面来看，鱼类种群在整个模拟过程中总体呈下降趋势，但下降的速度和稳定性因渔民的行为策略而异，并受到其贪婪因子的显著影响。当高贪婪因子的渔民主导捕捞活动时，鱼群数量下降更为迅速，反映出更具破坏性的资源开发模式。

引入如贪婪因子等行为参数，有助于对渔民的决策过程进行更为细致和现实的刻画(Wilen等，2002；Fulton等，2011)。通过调整这些参数，能够模拟从保守到激进的多种捕鱼行为，从而深入理解不同类型的行为如何影响公共资源的可持续性。研究表明，越激进、贪婪的捕捞行为往往导致资源更快枯竭，而保守、风险规避型行为则更有利于维持资源的长期稳定。将行为特征纳入资源管理模型，不仅有助于预测不同类型的资源使用者如何与“公地资源”互动，也为设计有效的可持续性干预措施提供了更全面的理论和实践视角。此外，通过整合行为、经济和偏好等社会数据，可以借助社会大数据同化对贪婪因子等行为参数进行更为精确的估计与校准(Ban等，2019；Li X等，2023)。

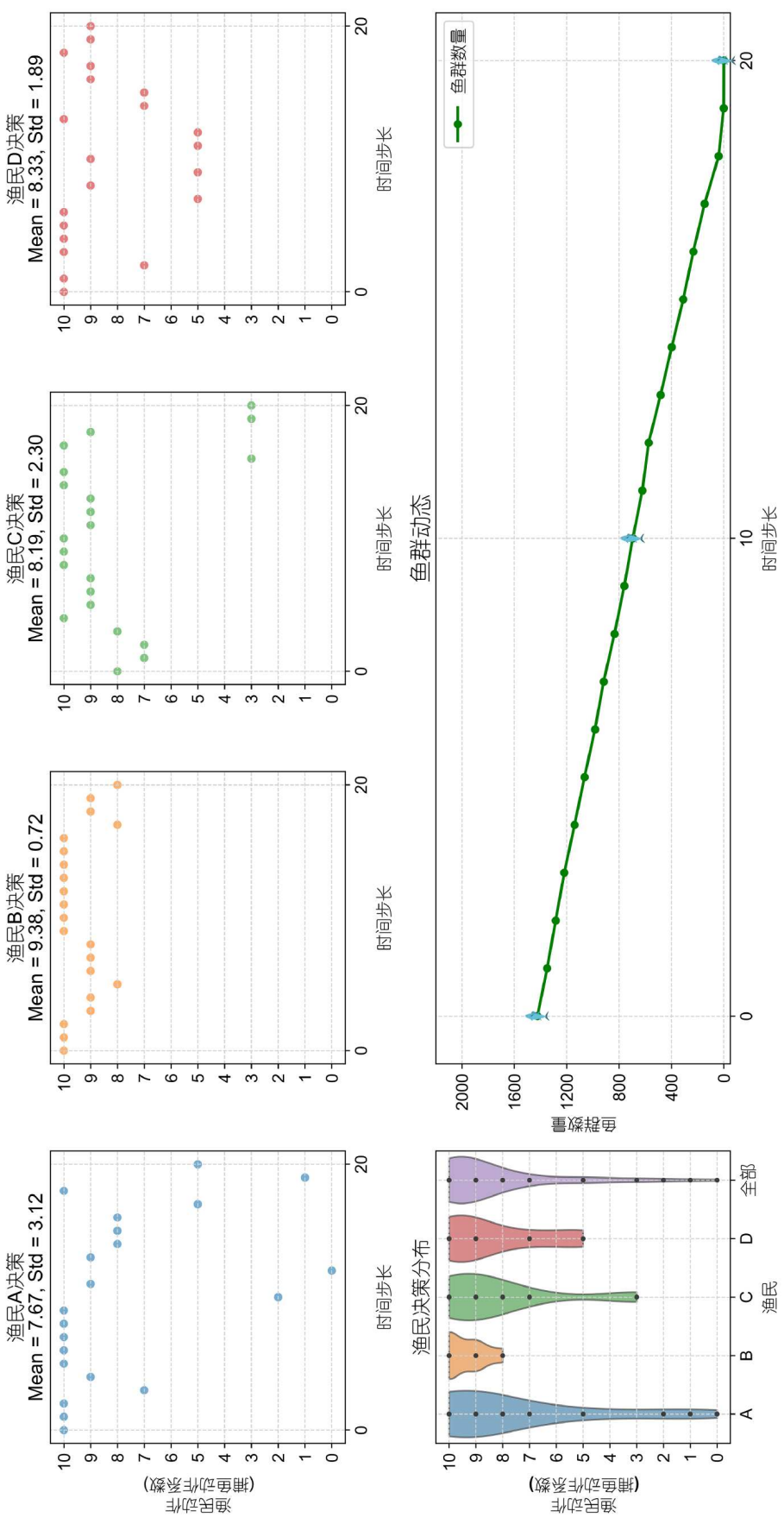


图 4　经济导向捕鱼行为下“公地悲剧”的涌现。该图展示了在以经济收益最大化为唯一目标的情景下，个体理性行为如何在集体层面导致资源崩溃。顶部：各渔民持续采取高强度捕鱼策略（Mean>7.5），行为变化较小，反映出所有渔民均专注于短期经济回报。左下：小提琴图揭示了所有渔民的行动分布高度集中于动作空间的上限区域，整体分布狭窄，表明其行为高度一致且趋于激进。右下：鱼群数量在20个时间步长内从初始的1600单位迅速下降至接近零，系统演化结果重现了经典的“公地悲剧”现象

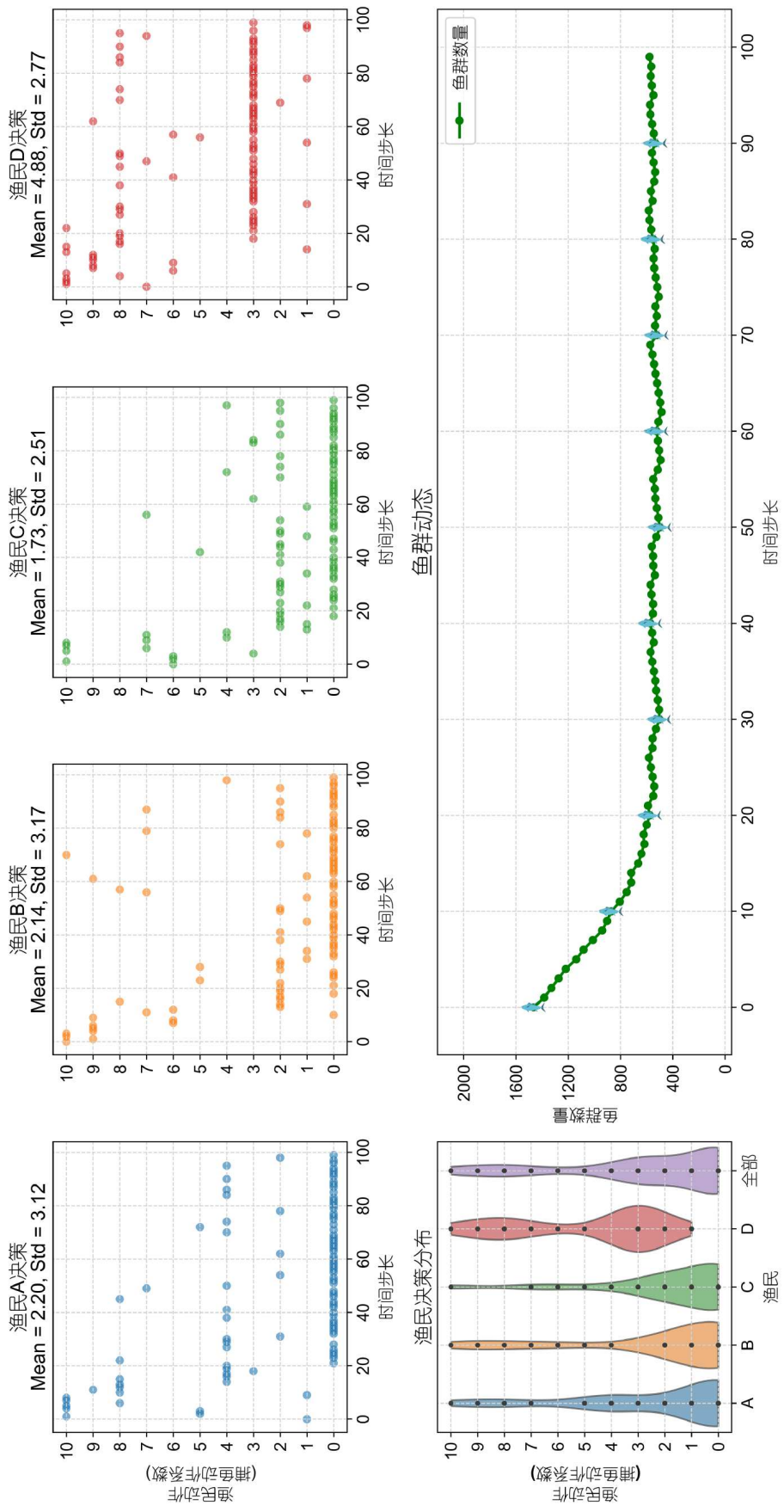


图 5 可持续性惩罚机制下可持续捕鱼模式的涌现。模拟结果展示了捕鱼策略向资源保护型策略的转变过程。顶部：各渔民表现出中等强度的捕捞行为，平均值介于1.73–4.88之间，并展现出显著的行为适应性。其中，渔民A、B和C保持较低的捕捞水平(Mean≤2.20)，而渔民D显示略高但稳定的捕捞强度(Mean=4.88)。左下：小提琴图揭示了渔民之间存在行为异质性，但整体趋向于保守的低强度捕捞策略，行动值主要集中在1–5的区间。右下：鱼群数量在初期从初始的1600单位有所下降，随后稳定在约500单位

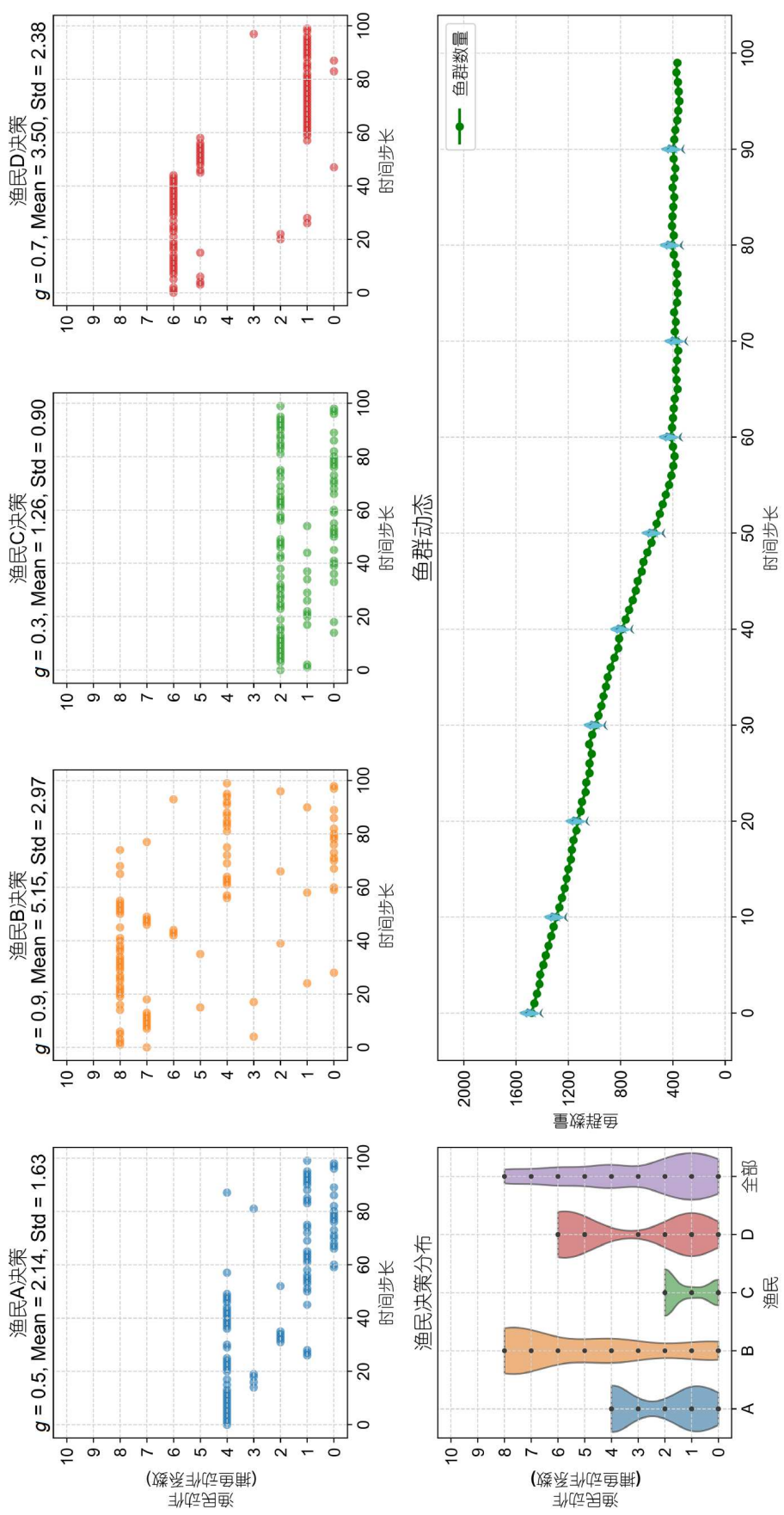


图 6　行为参数(贪婪因子g)对捕鱼行为和资源动态的影响。模拟结果展示了不同行为偏好下的渔民如何在个体策略与集体资源演化中产生差异化结果。顶部：渔民的捕鱼行为与贪婪因子g高度一致——低g值对应保守型行为(渔民C，g=0.3，Mean=1.26)，中等g值对应适度行为(渔民A和D，g≈0.6，Means：2.14–3.50)，而较高g值则对应激进、高强度的捕捞行为(渔民B，g=0.9，Mean=5.15)。左下：小提琴图清晰反映出渔民行为策略与贪婪因子之间的典型对应关系。右下：鱼群数量从1600单位逐渐下降并稳定在400单位左右

4.4.3 合作机制对渔民决策的影响

引入合作机制为促进渔业资源可持续发展提供了另一种有价值的视角。在该情景中，通过在奖励结构中引入合作因子，激励渔民在追求个体收益的同时，也关注整个群体的集体捕捞绩效。本案例中，合作因子δ_c被设定为0.6。

如图7所示，合作机制的引入显著改变了渔民的行为模式。与此前的情景相比，渔民的捕捞行为变得更加一致，行动变异性显著降低，极端行为明显减少。即使是具有较高贪婪因子(如渔民B，g=0.9)的个体，在合作奖励结构下也表现出更为稳定和适度的捕捞行为，不再展现出以往的激进倾向。小提琴图进一步揭示了捕捞强度分布的集中趋势，表明合作机制有效提升了渔民间的协调性与行为一致性。

从系统层面来看，鱼类种群动态也呈现出显著改善。在整个模拟过程中，鱼群数量基本维持在约1500单位，仅出现轻微波动，远高于未引入合作机制时的稳定水平。这一结果表明，合作机制能够有效减缓资源衰退，将鱼群维持在一个更高、更可持续的状态。通过将群体绩效纳入奖励结构，合作机制增强了渔民的集体责任意识，促使他们更倾向于做出有利于资源长期可持续性的决策。这一发现表明，设计合理的激励结构，能够引导资源使用者从个体理性向集体理性转变，从而缓解“公地悲剧”效应。

传统的资源管理模型通常假设个体行为完全由自身利益驱动，导致诸如“公地悲剧”等资源过度利用的结果。然而，本研究表明，通过在奖励结构中引入合作机制，可以有效引导个体行为与资源可持续性的整体目标相一致。在涉及公共资源的情境中，这种机制具有特别重要的意义，因为在此类环境中，每位参与者的决策都会直接或间接地影响其他相关方的福祉(Bendor和Scheffran，2018)。通过激励合作行为，政策制定者能够显著降低资源被过度开发的风险，从而构建更具韧性和可持续性的资源管理体系。

5 讨论和结论

人类世时代标志着一个关键的历史转折点，人类活动作为主导驱动力，以前所未有的速度推动地球自然系统发生远离平衡状态的深刻转变。这些人为扰动引发了人类与自然系统之间复杂的协同演化动态，其特征包括非线性反馈机制、路径依赖以及系统状态的分化演变。然而，当前主流的建模框架在刻画此类复杂动态方面仍存在显著不足，现有的数学模型无法有效描述系统内部的反馈机制、状态转换和临界点，从而限制了对复合型危机情境下系统行为的预测与应对能力。

为应对上述挑战，本研究将人地系统定义视为复杂适应系统，并将其映射到由四元组构成的马尔可夫决策过程，包括环境状态(涵盖自然-社会-生态系统)、主体行动(从个体到群体)、状态转移(系统演化)和奖励(行动回报)。基于这一框架，本研究的主要贡献和创新性体现在三个方面。

(1)构建人地系统的通用建模与决策框架：提出了一个通用的多主体深度强化学习框架，该框架将马尔可夫决策过程、多主体建模和深度强化学习有机结合，为解决人地系统建模和决策难题提供了有力工具。

(2)对经典问题的建模验证与机制揭示：通过经典的“公地悲剧”问题展示了该框架的有效性，深入分析了个体行为偏好、奖励结构设计及合作机制如何共同影响资源可持续性，揭示了复杂个体行为和环境反馈如何导致了资源枯竭的涌现。

(3)治理机制的量化评估：研究结果还对不同治理策略进行了定量评估。结果表明，采用自适应决策机制和合作激励策略，能够显著提升资源利用的可持续性，其效果远优于单纯追求短期经济利益的策略。这突出了在资源管理中协调个体行为与集体目标的重要性。

尽管本研究为人地系统的建模和决策提供了有价值的见解，但仍只是对构建人地系统通用框架的初步探索。因此，有必要系统总结当前存在的局限性，并为未来研究指明方向。首先，本研究的发现受限于缺乏大规模实验支持，制约了在现实环境中探索复杂互动机制和应对“维数灾难”等高维决策问题的能力。其次，尽管本研究中的环境模型为理论探索提供了基础，但其在反映真实生态系统的复杂性与动态变异性方面仍存在不足。虽然模型中引入了逻辑斯蒂增长函数等机制以部分刻画生态动态，但诸如气候变化、栖息地演变以及其他复杂生态过程尚未被纳入建模框架。这在一定程度上削弱了模型对现实人地系统进行有效模拟与调控的能力。此外，模型在环境可观测性与不确定性建模方面仍有待加强，尽管已有部分研究在该方向上进行了初步探索(Lane，1989；van Dijk等，2014)。再次，当前研究仅聚焦于有限类型的主体(如渔民)和较少的主体数量(如四个)，未能充分考虑现实世界中更为多元的利益相关者结构，如政府、机构和社区组织等。尽管已有研究探讨了个体层面的决策机制(Dreyfus-León，1999；Wilen等，2002)，近期研究也尝试引入层级结构(Nicolas和Paul-Antoine，2020，2021)，但要全面刻画不同利益相关方之间的目标差异、权力结构与利益博弈仍面临显著挑战。最后，主体间的决策过程与互动机制在本研究中被相对简化处理。

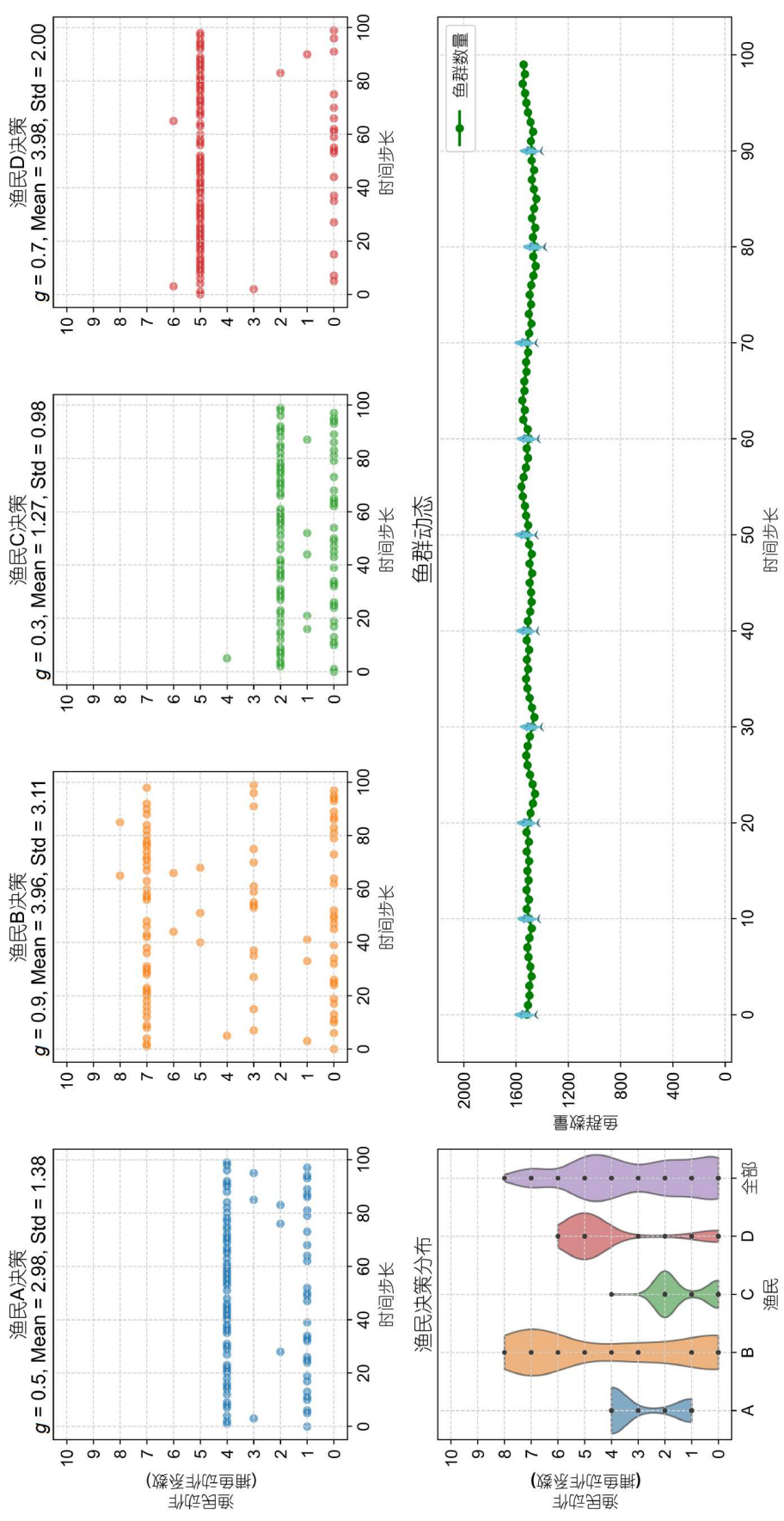


图 7　合作机制对捕鱼行为和资源动态的影响。模拟结果显示了如何通过合作在个体倾向多样化的情况下实现资源的稳定管理。顶部：渔民展现出与其贪婪因子g相一致但趋于收敛的行为模式——渔民C(g=0.3)保持保守捕鱼(Mean=1.27)，而其他渔民虽表现出中等到较高的捕捞活动(Means：2.98–3.98)，但其行为波动显著减小，整体趋于中等强度范围。左下：小提琴图展示了在合作机制下渔民的捕捞强度分布呈现出更为集中和一致的趋势。右下：最值得注意的是，鱼类种群在整个模拟期间保持在1500单位左右的高稳定水平

主要考虑合作和竞争等基础策略，风险承担、信任建立、社会学习等更细致的行为模式未被充分纳入。未来研究应进一步整合学习能力、适应性行为和动态博弈机制，以更好地反映现实主体间的复杂互动(Nicolas和Paul-Antoine，2020)。

未来的研究将进一步完善这些方面，并拓展当前框架以支持现实世界中的大规模应用，从而提升对复杂耦合系统的建模与决策能力。为此，我们将从以下三个维度融合机器学习与数据驱动方法：(1)先进的环境建模技术：通过数据同化技术将过程模型与卫星观测、物联网传感器网络以及社会经济调查等多元数据进行动态融合(Li等，2024)。这种混合建模方法将有效解决传统模型中存在的时空不匹配问题。同时，借助机器学习替代模型(如物理信息神经网络)来加速对非线性反馈机制和随机扰动的模拟过程(丁琳等，2025；Li和Liu，2025)。(2)精细化的主体行为建模：构建理论-数据双驱动的主体认知机制，使不同利益相关者能够模拟在资源枯竭或政策变化情境下的动态策略调整行为。该方法将超越传统的合作与竞争范式，引入风险敏感型决策、制度性谈判等更贴近现实的行为机制(Li和Guo，2025)。(3)多层次的决策支持架构：建立跨越微观(个体)、中观(区域)和宏观(国家)的多层次决策支持系统，配备激励和约束机制，以协调跨尺度决策与系统目标之间的关系(Li等，2018，2022；Ge等，2022)。该架构将帮助政策制定者开展多维不确定性评估(如生态保护政策、碳市场干预)，使其在政策实施前识别因果效应而非仅仅只是相关性(苏建宾等，2023)。最终实现人类世背景下人地系统动态耦合理论机制与可操作治理策略的深度融合与协同转化。

参考文献

陈发虎, 李新, 吴绍洪, 等. 2021. 中国地理科学学科体系浅析. 地理学报, 76: 2069–2073

丁琳, 摆玉龙, 郑东海, 等. 2025. 融合稀疏观测与改进回声状态网络的混沌气候系统预测. 中国科学: 地球科学, 55: 2425–2439

刘彦随, 刘亚群, 欧聪. 2024. 现代人地系统科学认知与探测方法. 科学通报, 69: 447–463

钱学森. 1991. 谈地理科学的内容及研究方法(在1991年4月6日中国地理学会“地理科学”讨论会上的发言). 地理学报, 46: 257–265

苏建宾, 陈都鑫, 郑东海, 等. 2023. 追索为什么? 地球系统科学中的因果推理. 中国科学: 地球科学, 53: 2199–2216

原世伟, 李新, 杜二虎. 2021. 多主体建模在水资源管理中的应用: 进展与展望. 地球科学进展, 36: 899–910

岳天祥, 吴晨辰, 史文娇, 等. 2024. 人类-自然系统耦合模型研究动态. 中国科学: 地球科学, 54: 3668–3673

Ale Ebrahim Dehkordi M, Lechner J, Ghorbani A, et al. 2023. Using machine learning for agent specifications in agent-based models and simulations: A critical review and guidelines. JASSS, 26: 9

An L. 2012. Modeling human decisions in coupled human and natural systems: Review of agent-based models. Ecol Model, 229: 25–36

Arulkumaran K, Deisenroth M P, Brundage M, et al. 2017. Deep reinforcement learning: A brief survey. IEEE Signal Process Mag, 34: 26–38

Ban N C, Gurney G G, Marshall N A, et al. 2019. Well-being outcomes of marine protected areas. Nat Sustain, 2: 524–532

Bankes S C. 2002. Agent-based modeling: A revolution? Proc Natl Acad Sci USA, 99: 7199–7200

Bellman R. 1954. The theory of dynamic programming. Bull Amer Math Soc, 60: 503–515

Bellman R. 1957a. A Markovian decision process. J Math Mech, 6: 679–684

Bellman R. 1957b. Dynamic Programming. Princeton: Princeton University Press

Bellman R. 1966. Dynamic programming. Science, 153: 34–37

Bendor T, Scheffran J. 2018. Agent-Based Modeling of Environmental Conflict and Cooperation. Boca Raton: CRC Press

Bengio Y, Courville A, Vincent P. 2013. Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell, 35: 1798–1828

Berry C J. 2007. Hume’s universalism: The science of man and the anthropological point of view. Br J Hist Philos, 15: 535–550

Bodin Ö, Chen H. 2023. A network perspective of human-nature interactions in dynamic and fast-changing landscapes. Natl Sci Rev, 10: nwad019

Bonabeau E. 2002. Agent-based modeling: Methods and techniques for simulating human systems. Proc Natl Acad Sci USA, 99: 7280–7287

Bone C, Dragićević S. 2010. Simulation and validation of a reinforcement learning agent-based model for multi-stakeholder forest management. Comput Environ Urban Syst, 34: 162–174

Costello C, Cao L, Gelcich S, et al. 2020. The future of food from the sea. Nature, 588: 95–100

Dreyfus-León M. 1999. Individual-based modelling of fishermen search behaviour with neural networks and reinforcement learning. Ecol Model, 120: 287–297

Epstein J M, Axtell R. 1996. Growing Artificial Societies: Social Science from the Bottom Up. Washington: Brookings Institution Press

Farahbakhsh I, Bauch C T, Anand M. 2022. Modelling coupled human-environment complexity for the future of the biosphere: Strengths, gaps and promising directions. Phil Trans R Soc B, 377: 20210382

Farmer J D, Foley D. 2009. The economy needs agent-based modelling. Nature, 460: 685–686

Fulton E A, Smith A D M, Smith D C, et al. 2011. Human behaviour: The key source of uncertainty in fisheries management. Fish Fisher, 12: 2–17

Ge Y, Li X, Cheng G, et al. 2022. What dominates sustainability in endorheic regions? Sci Bull, 67: 1636–1640

Grimm V, Revilla E, Berger U, et al. 2005. Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. Science, 310: 987–991

Hardin G. 1968. The tragedy of the commons. Science, 162: 1243–1248

Hernandez-Leal P, Kartal B, Taylor M E. 2019. A survey and critique of multiagent deep reinforcement learning. Auton Agent Multi-Agent Syst, 33: 750–797

Holland J H. 1992. Complex Adaptive Systems. Daedalus, 121: 17–30

Holland J H. 1995. Hidden Order: How Adaptation Builds Complexity. Reading: Addison-Wesley

Holland J H. 2002. Complex adaptive systems and spontaneous emergence. In: Curzio A Q, Fortis M, eds. Complexity and Industrial Clusters: Dynamics and Models in Theory and Practice. Heidelberg: Physica-Verlag. 25–34

Holling C S. 1973. Resilience and stability of ecological systems. Annu Rev Ecol Syst, 4: 1–23

Hume D. 1739. A Treatise of Human Nature. Oxford: Clarendon Press

Ju J, Kurniawati H, Kroese D, et al. 2025. Model-based offline reinforcement learning for sustainable fishery management. Expert Syst, 42: e13324
Kedir N S, Somi S, Fayek A R, et al. 2022. Hybridization of reinforcement learning and agent-based modeling to optimize construction planning and scheduling. Automat Constr, 142: 104498
Lane D E. 1989. A partially observable model of decision making by fishermen. Oper Res, 37: 240–254
Lansing J S. 2003. Complex adaptive systems. Annu Rev Anthropol, 32: 183–204
Lee K, Rucker M, Scherer W T, et al. 2017. Agent-based model construction using inverse reinforcement learning. In: Proceedings of the 2017 Winter Simulation Conference (WSC). Las Vegas: IEEE. 1264–1275
Lenton T M, Held H, Kriegler E, et al. 2008. Tipping elements in the Earth's climate system. Proc Natl Acad Sci USA, 105: 1786–1793
Levin S, Xepapadeas T, Crépin A S, et al. 2013. Social-ecological systems as complex adaptive systems: Modeling and policy implications. Envir Dev Econ, 18: 111–132
Li X, Guo Y. 2025. Paradigm shifts from data-intensive science to robot scientists. Sci Bull, 70: 14–18
Li X, Liu F. 2025. A mathematical comparison of data assimilation and machine learning in earth system state estimation from a Bayesian inference viewpoint. Inf Geogr, 1: 100001
Li X, Cheng G, Lin H, et al. 2018. Watershed system model: The essentials to model complex human-nature system at the river basin scale. J Geophys Res-Atmos, 123: 3019–3034
Li X, Cheng G, Fu B, et al. 2022. Linking critical zone with watershed science: The example of the Heihe River basin. Earths Future, 10: e2022EF002966
Li X, Feng M, Ran Y, et al. 2023. Big Data in Earth system science and progress towards a digital twin. Nat Rev Earth Environ, 4: 319–332
Li X, Liu F, Ma C, et al. 2024. Land data assimilation: Harmonizing theory and data in land surface process studies. Rev Geophys, 62: e2022RG000801
Li Y, Sang S, Mote S, et al. 2023. Challenges and opportunities for modeling coupled human and natural systems. Natl Sci Rev, 10: nwad054
Liu J, Dietz T, Carpenter S R, et al. 2007. Complexity of coupled human and natural systems. Science, 317: 1513–1516
Lorenz E N. 1963. Deterministic nonperiodic flow. J Atmos Sci, 20: 130–141
May R M. 1976. Simple mathematical models with very complicated dynamics. Nature, 261: 459–467
Meadows D H, Meadows D L, Randers J, et al. 1972. The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind. New York: Universe Books
Memarzadeh M, Britten G L, Worm B, et al. 2019. Rebuilding global fisheries under uncertainty. Proc Natl Acad Sci USA, 116: 15985–15990
Minsky M. 1961. Steps toward artificial intelligence. Proc IRE, 49: 8–30
Mnih V, Kavukcuoglu K, Silver D, et al. 2015. Human-level control through deep reinforcement learning. Nature, 518: 529–533
Nicolas P G, Paul-Antoine B. 2020. Deep hierarchical reinforcement learning in a Markov game applied to fishery management decision making. In: Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI). Canberra: IEEE. 1371–1378
Nicolas P G, Paul-Antoine B. 2021. Deep hierarchical reinforcement learning to manage the trade-off between sustainability and profitability in common pool resources systems. In: Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN). Shenzhen: IEEE. 1–7
Ostrom E. 1990. Governing the Commons: The Evolution of Institutions for Collective Action. Cambridge: Cambridge university press
Puterman M L. 1990. Markov decision processes. In: Linetsky V, Birge J, eds. Handbooks in Operations Research and Management Science. Amsterdam: Elsevier Science. 331–434
Reed P M, Hadjimichael A, Moss R H, et al. 2022. Multisector dynamics: Advancing the science of complex adaptive human-Earth systems. Earths Future, 10: e2021EF002621
Rispoli G. 2020. Genealogies of Earth System thinking. Nat Rev Earth Environ, 1: 4–5
Rockström J, Steffen W, Noone K, et al. 2009. A safe operating space for humanity. Nature, 461: 472–475
Rudd M B, Branch T A. 2017. Does unreported catch lead to overfishing? Fish Fisher, 18: 313–323
Schulze J, Müller B, Groeneveld J, et al. 2017. Agent-based modelling of social-ecological systems: Achievements, challenges, and a way forward. Int J Commons, 20: 8
Sert E, Bar-Yam Y, Morales A J. 2020. Segregation dynamics with reinforcement learning and agent based modeling. Sci Rep, 10: 11771
Sethi G, Costello C, Fisher A, et al. 2005. Fishery management under multiple uncertainty. J Environ Econ Manage, 50: 300–318
Silver D, Schrittwieser J, Simonyan K, et al. 2017. Mastering the game of Go without human knowledge. Nature, 550: 354–359
Silver D, Hubert T, Schrittwieser J, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science, 362: 1140–1144
Skinner B F. 1965. Science and Human Behavior. New York: Simon and Schuster
Smith A. 1776. An Inquiry into the Nature and Causes of the Wealth of Nations. London: W Strahan and T Cadell
Steffen W, Richardson K, Rockström J, et al. 2020. The emergence and evolution of Earth System Science. Nat Rev Earth Environ, 1: 54–63
Sutton R S. 1984. Temporal credit assignment in reinforcement learning. Dissertation for Doctoral Degree. Amherst: University of Massachusetts Amherst
Sutton R S, Barto A G. 2018. Reinforcement Learning: An Introduction. Cambridge: The MIT Press
The Nobel Committee for Physics. 2021. Scientific Background on the Nobel Prize in Physics 2021
van Dijk D, Hendrix E M T, Haijema R, et al. 2014. On solving a bi-level stochastic dynamic programming model for analyzing fisheries policies: Fishermen behavior and optimal fish quota. Ecol Model, 272: 68–75
Von Neumann J. 1928. Zur theorie der gesellschaftsspiele. Math Ann, 100: 295–320
Von Neumann J, Morgenstern O. 1944. Theory of games and economic behavior. Princeton: Princeton University Press
Wang X, Wang S, Liang X, et al. 2022. Deep reinforcement learning: A survey. IEEE Trans Neural Netw Learn Syst, 35: 5064–5078
Watkins C. 1989. Learning from delayed rewards. Dissertation for Doctoral Degree. Cambridge: King's College, University of Cambridge
Werbos P. 1977. Advanced forecasting methods for global crisis warning and models of intelligence. Gen Syst Yearb, XXII: 25–38
Wilen J E, Smith M D, Lockwood D, et al. 2002. Avoiding surprises: Incorporating fisherman behavior into management models. Bull Mar Sci, 70: 553–575
Williams B K. 2011. Resolving structural uncertainty in natural resources management using POMDP approaches. Ecol Model, 222: 1092–1102
Yuan S, Li X, Du E. 2021. Effects of farmers' behavioral characteristics on crop choices and responses to water management policies. Agric Water Manage, 247: 106693

（责任编委：汤秋鸿）